



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### A Parsed Linguistic Atlas of Early Middle English

**Citation for published version:**

Truswell, R, Alcorn, R & Donaldson, J 2016, 'A Parsed Linguistic Atlas of Early Middle English', Paper presented at Angus McIntosh Centre for Historical Linguistics Symposium, Edinburgh, United Kingdom, 9/06/16 - 10/06/16.

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Other version

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# A Parsed Linguistic Atlas of Early Middle English

Rob Truswell, Rhona Alcorn, and James Donaldson  
University of Edinburgh

`rob.truswell@ed.ac.uk, r.alcorn@ed.ac.uk,`  
`james.donaldson@ed.c.uk`

AMC symposium, 10/6/16

# Outline

- ▶ An initial report on a new, BA/Leverhulme-funded, corpus project.
- ▶ Goal: use data from the Linguistic Atlas of Early Middle English to plug a gap in the Penn Parsed Corpora of Historical English.
- ▶ Agenda:
  1. Beginner's guide to PPCHE
  2. Beginner's guide to LAEME
  3. Plans for the project
  4. Beyond this project

## Section 1

PPCHE

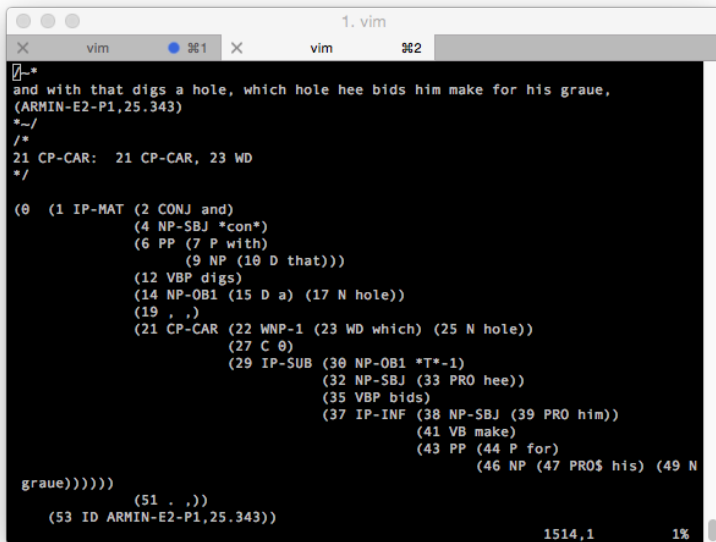
# The Penn historical corpus format

- ▶ Family of parsed historical corpora covering English (7.7m words, –1914) and several other Germanic and Romance languages.
- ▶ Annotated with detailed labelled brackets.
  - ▶ Explicit annotation of grammatical function.
  - ▶ Explicit representation of many cases of “movement”, extraposition, etc.
- ▶ The PPCHE philosophy:
  - ▶ no claim to being an accurate theory of syntax.
  - ▶ aim to include useful information about constituency in a way that can be consistently implemented.
  - ▶ So (almost) no VP nodes, default high attachment, etc.
- ▶ Major virtue: very easy to query information about phrasal syntax.
- ▶ Very easy to make (some kind of) crosslinguistic comparison.

## Sample query

[illegible]

## Sample query: sample output



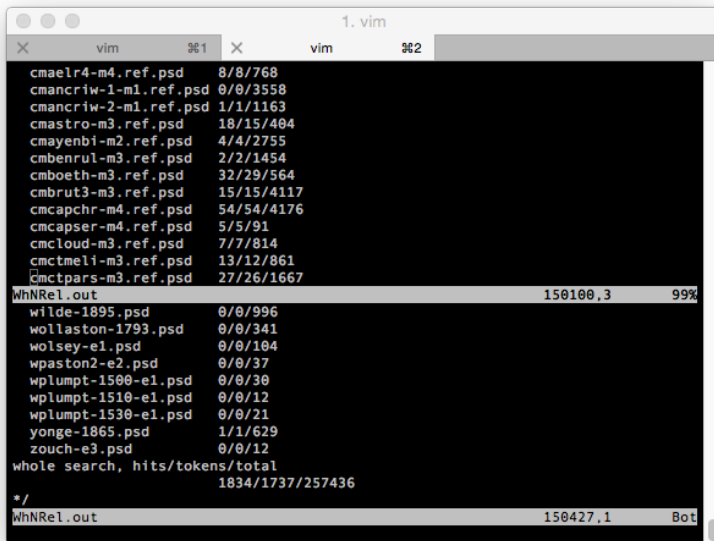
The screenshot shows a vim editor window titled "1. vim". The editor has two tabs: "vim" (active) and "vim" (inactive). The active tab shows a sample query and its output. The query is a sentence: "and with that digs a hole, which hole hee bids him make for his graue, (ARMIN-E2-P1,25.343)". The output is a parse tree for the sentence, showing the hierarchical structure of the sentence. The parse tree is rooted at (0) (1 IP-MAT (2 CONJ and) (4 NP-SBJ \*con\*) (6 PP (7 P with) (9 NP (10 D that))) (12 VBP digs) (14 NP-OB1 (15 D a) (17 N hole)) (19 , ,) (21 CP-CAR (22 WNP-1 (23 WD which) (25 N hole)) (27 C 0) (29 IP-SUB (30 NP-OB1 \*T\*-1) (32 NP-SBJ (33 PRO hee)) (35 VBP bids) (37 IP-INF (38 NP-SBJ (39 PRO him)) (41 VB make) (43 PP (44 P for) (46 NP (47 PRO\$ his) (49 N graue)))))). The output is displayed in a monospaced font on a black background. The status bar at the bottom right shows "1514,1" and "1%".

```
1. vim
vim  %1 vim  %2
~*
and with that digs a hole, which hole hee bids him make for his graue,
(ARMIN-E2-P1,25.343)
*~/
/*
21 CP-CAR:  21 CP-CAR, 23 WD
*/

(0  (1 IP-MAT (2 CONJ and)
      (4 NP-SBJ *con*)
      (6 PP (7 P with)
            (9 NP (10 D that)))
      (12 VBP digs)
      (14 NP-OB1 (15 D a) (17 N hole))
      (19 , ,)
      (21 CP-CAR (22 WNP-1 (23 WD which) (25 N hole))
                (27 C 0)
                (29 IP-SUB (30 NP-OB1 *T*-1)
                          (32 NP-SBJ (33 PRO hee))
                          (35 VBP bids)
                          (37 IP-INF (38 NP-SBJ (39 PRO him))
                                    (41 VB make)
                                    (43 PP (44 P for)
                                            (46 NP (47 PRO$ his) (49 N
graue))))))
      (51 . ,))
      (53 ID ARMIN-E2-P1,25.343))

1514,1 1%
```

# Sample query: counts



The screenshot shows a vim terminal window titled "1. vim". The terminal displays a list of files and their counts, followed by a summary line. The files are listed in two columns, with the first column containing the file name and the second column containing the count. The summary line is highlighted in gray. The terminal also shows a status bar at the bottom with the text "WhNRel.out", "150427,1", and "Bot".

```
vim %1 vim %2
cmaelr4-m4.ref.psd 8/8/768
cmancriw-1-m1.ref.psd 0/0/3558
cmancriw-2-m1.ref.psd 1/1/1163
cmastro-m3.ref.psd 18/15/404
cmayenbi-m2.ref.psd 4/4/2755
cmbenrul-m3.ref.psd 2/2/1454
cmboeth-m3.ref.psd 32/29/564
cmbrut3-m3.ref.psd 15/15/4117
cmcapchr-m4.ref.psd 54/54/4176
cmcapser-m4.ref.psd 5/5/91
cmcloud-m3.ref.psd 7/7/814
cmctmeli-m3.ref.psd 13/12/861
cmctpars-m3.ref.psd 27/26/1667
WhNRel.out 150100,3 99%
wilde-1895.psd 0/0/996
wollaston-1793.psd 0/0/341
wolsey-e1.psd 0/0/104
wpaston2-e2.psd 0/0/37
wplumtp-1500-e1.psd 0/0/30
wplumtp-1510-e1.psd 0/0/12
wplumtp-1530-e1.psd 0/0/21
yonge-1865.psd 1/1/629
zouch-e3.psd 0/0/12
whole search, hits/tokens/total
*/ 1834/1737/257436
WhNRel.out 150427,1 Bot
```



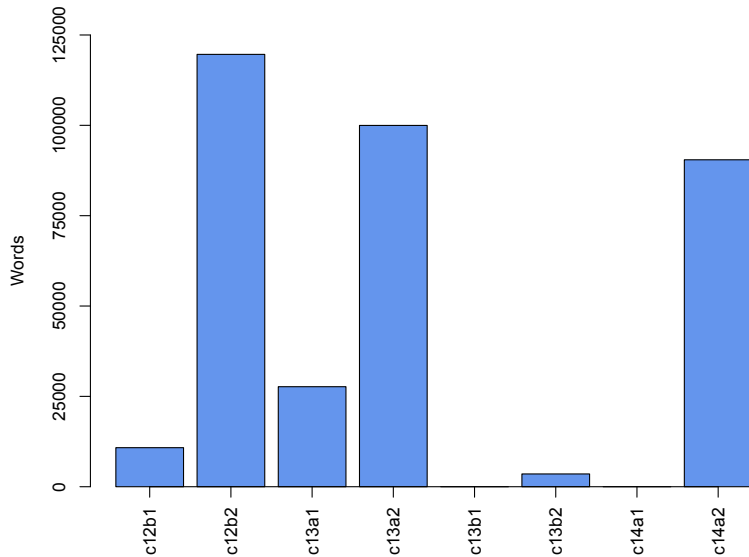
## PPCME2 weaknesses

- ▶ Built from published editions, not manuscripts.
- ▶ Limited metadata.
- ▶ Not lemmatized.
- ▶ Significant data gap c.1250–1340.

## The data gap: PPCME2, 1150–1350

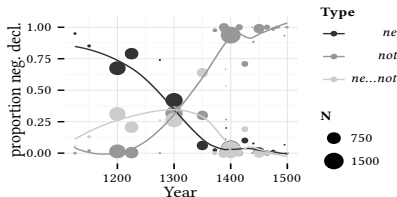
Filename	Title	Date	Words
cmkenth	Kentish Homilies	c12a2–b1	4048
cmpeterb	Peterborough Chronicle	c.1131, c.1154	6757
cmlambx1	Lambeth Homilies	c12b2	20752
cmtrinit	Trinity Homilies	c12b2	41844
cmorm	Ormulum	c12b2	50579
cmlamb1	Lambeth Homilies	c12b2	6459
cmvices1	Vices and Virtues	c13a1	27677
cmsawles	Sawles Warde	c13a2	4111
cmhali	Hali Meiðhad	c13a2	8495
cmkathe	St. Katherine	c13a2	8699
cmjulia	St. Juliana	c13a2	6810
cmmarga	St. Margaret	c13a2	8069
cmancriw	Ancrene Riwle	c13a2	63790
cmkentse	Kentish Sermons	c13b2?	3534
cmayenbi	Ayenbite of Inwyte	1340	45944
cmearlps	Earliest Prose Psalter	c.1350	44521

## The data gap: PPCME2, 1150–1350

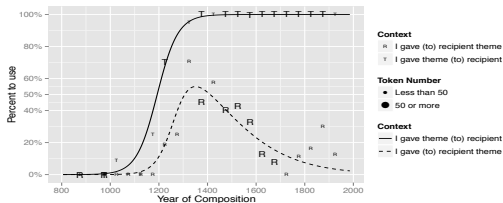


# The data gap in action

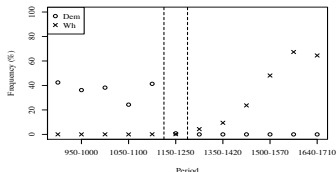
Ecay & Tamminga  
(2016): Expression  
of negation



Bacovcin (2016):  
Double object word  
order



Gisborne & Truswell  
(2016): Argument  
relatives



## Help is at hand

- ▶ Luckily, the Linguistic Atlas of Early Middle English (Laing 2013) complements PPCME2 remarkably well.
  - ▶ Built from manuscripts.
  - ▶ Extensive metadata.
  - ▶ Lemmatized.
  - ▶ Significant amounts of material from 1250–1325.
- ▶ A major motivation for constructing PLAEME is that LAEME complements PPCME2 so well.
- ▶ By transferring features from LAEME, we also have the opportunity to expand on the PPCHE format.

## Section 2

LAEME

## LAEME

- ▶ Compilation began early '90s; published online 2008
- ▶ Specimens of written English 1150 – 1325:
  - Documents, e.g. Chertsey Cartulary
  - Prose, e.g. *Vices and Virtues*
  - Poetry, e.g. *Poema Morale*
  - Lyrics, e.g. *Love Song of Our Lady*
- ▶ c.650,000 words (shortest 18; longest 30,237)

## Transcription

- ▶ Upper case for 'plain text' ms letters
- ▶ \* for capitals
- ▶ Lower case for special characters:
  - ▶ y = þ, d = ð, w = p, ae = æ, g = ġ, z = 3

MS: <For it is caldore ner þe se> (*The Infancy of Christ*)

LAEME: \*FOR IT IS CALDORE NER yE SE



## Tagging

- ▶ Most morphemes (incl. inflections & affixes) are tagged
- ▶ \$lexel/grammel\_form, e.g. \$for/cj\_\*FOR
  - ▶ lexel: semantic identifier
  - ▶ grammel: grammatical identifier
- ▶ All tags can be translated into lemmas
- ▶ Word-level tags arranged vertically in tagged text files

```
$for/cj_*FOR  
$/P13NI_IT  
$be/vps13_IS  
$cold/aj-cpv_CALD+ORE  $-er/xs-cpv-aj_+ORE  
$near/pr_NER  
$/T<pr_yE  
$sea/n<pr_SE
```

## Syntactic clues: NPs

\$near/pr\_NER  
\$/**T**<pr\_**yE**  
\$sea/**n**<pr\_**SE**

- ▶  $\$/T(\dots) = \text{Definite determiner}$
- ▶  $\$/xxx/n(\dots) = \text{Noun}$
- ▶  $\$/T(\dots) + \$/xxx/n(\dots) = \text{NP}$

## Syntactic clues: NP function

\$near/pr\_NER  
\$/T<**pr**\_yE  
\$sea/n<**pr**\_SE

- ▶ <pr = 'in the scope of a preceding preposition'
- ▶ \$/T<pr + \$xxx/n<pr = NP object of P

## Syntactic clues: NP function

<To teche þe volk þe rizte lawe>  
'to teach the folk the right law'

\$to/im+C\_\*TO  
\$teach/vi-m\_TECH+E \$/vi-m\_+E  
\$/T**Oi**\_yE  
\$folk/n**Oi**\_VOLK  
\$/TOd\_yE  
\$right/ajOd\_RIzTE  
\$law/nOd\_LAWE

- ▶  $\$/T(\dots) + \$folk/n(\dots) = NP$
- ▶ Oi in grammel indicates indirect object function

## Syntactic clues: NP function

\$to/im+C\_\*TO  
\$teach/vi-m\_TECH+E \$/vi-m\_+E  
\$/TOi\_yE  
\$folk/nOi\_VOLK  
\$/T**Od**\_yE  
\$right/aj**Od**\_RIzTE  
\$law/n**Od**\_LAW

- ▶  $\$/T(\dots) + \$right/aj(\dots) + \$law/n(\dots) = NP$
- ▶ Od in grammel indicates direct object function

## Non-local dependency: discontinuous XPs

<... ðat ghe ne migte **him** bringen **on**>

'... which she may not prove **against him**'

- ▶ >pr = 'object of a following preposition'
- ▶ pr< = 'preposition with a preposed object'

```
$/RTIOd_dAT  
$/P13NF_GHE  
$/neg-v_NE  
$may/vpt13_MIGTE  
$/P13>prM_HIm  
$bring/vi_BRING+EN $/vi_+EN  
$on{p}/pr<{rh}_ON
```

## Non-local dependency: 'braced' negation

- ▶  $\geq$  points forward to a non-contiguous coordinating item
- ▶  $\leq$  points back to a non-contiguous coordinating item

<|pou **ne** sselt **naʒt** *consenti* ...>  
'thou shall not consent ...'

```
$/P12N_yOU  
$/neg-v $\geq$ _NE  
$shall/vps12_SSELT  
$not/neg-v $\leq$ _NAzT  
$consent/viFir_conSENT+I $/viFir_+I
```

## LAEME: Text selection

Prioritised texts:

1. 1250 - 1325
2. > 100 words
3. No pre-existing parsed version

Multiple versions of same text:

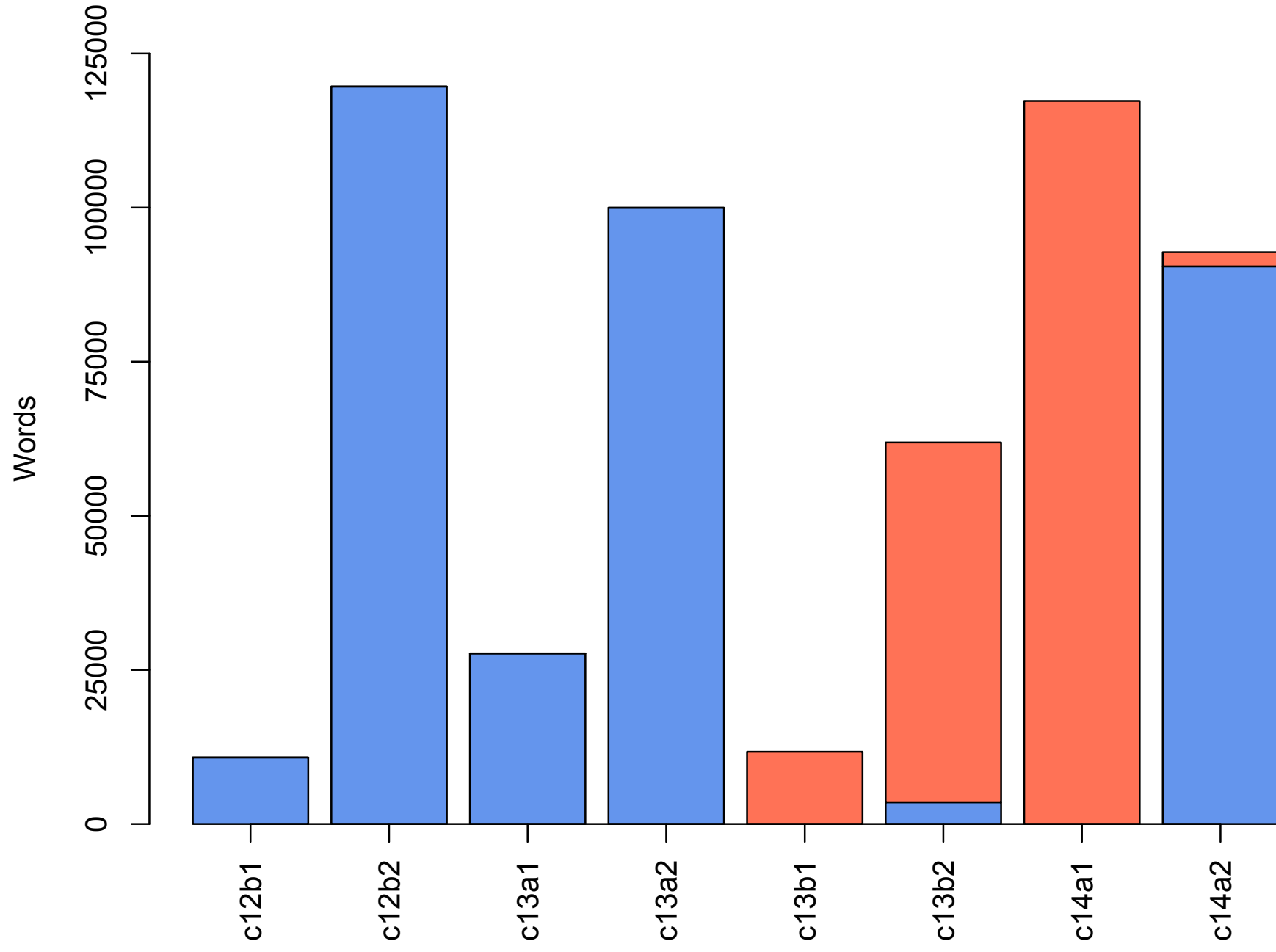
4. Choose on basis of date, length and county



## Our top 8

<b>Text</b>	<b>#</b>	<b>Date</b>	<b>Words</b>	<b>County</b>
<i>S. English Legendary</i>	286	ca.1310-20	30,237	Berks
Nthn Homily Coll'n	298	C14a	22,164	Yorks NR
<i>Havelok</i>	285	C14a1	17,089	Norfolk
<i>Cursor Mundi</i> , Hand A	297	C14a	15,107	Yorks ER
<i>Cursor Mundi</i> , Hand C	296	C14a	14,087	York
<i>Infancy of Christ</i>	283	ca.1300	12,489	Oxon
<i>Genesis and Exodus</i>	155	C14a1	12,467	Norfolk
<i>Life of Christ</i>	281	ca.1300	10,547	Oxon

These texts alone supplement PPCHE by 134,187 words



## Section 3

### The parsing process

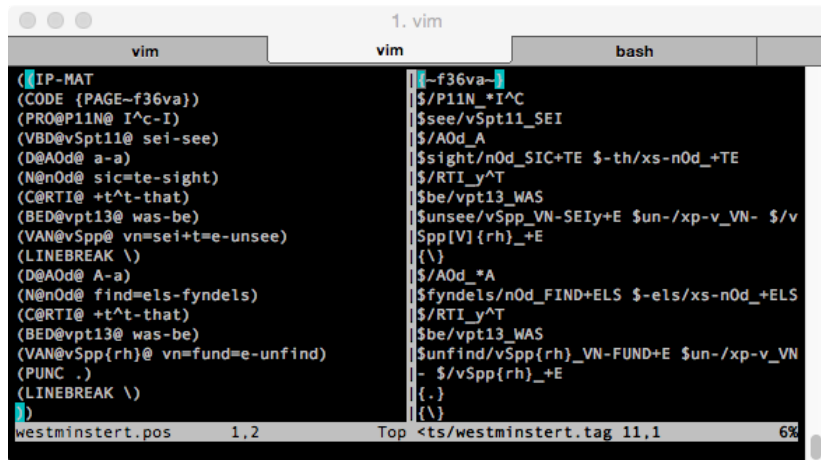
# Distinctive properties

- ▶ Unusual parsing approach:
    - ▶ LAEME grammels stand in a many-many relationship with PPCHE POS tags, so will not be directly visible in the output.
    - ▶ Nevertheless, LAEME grammels contain a large amount of syntactically relevant information:
      - ▶ Category
      - ▶ Grammatical function
      - ▶ Some meaning relations
      - ▶ Some nonlocal dependencies
- `$/neg-v([neither]<av>norC>nor>norC)_NE`  
(laud108bt.tag, line 1161)
- ▶ Plan in essence: project the LAEME grammels into labelled brackets, while also replacing them with PPCHE tags.
  - ▶ Allows for fairly accurate and very detailed automatic structure generation.

## Workflow: Format conversion

1. Take LAEME text
2. Store initial metadata as separate file.
3. For lines with textual material (initial character \$, ', ;):
  - 3.1 Find full word form, most informative grammel, etc.
  - 3.2 Segment that material into \$lexel/grammel\_form.
  - 3.3 Convert LAEME orthography into PPCHE orthography.
  - 3.4 Add lemmata for function words (no lexels in LAEME).
  - 3.5 Reformat as (POS@grammel@ form-lexel), where POS is a best-guess PPCHE equivalent of the LAEME grammel, with some lexel-by-lexel correction for e.g. different treatment of relativizers, subordinating conjunctions, etc.).
  - 3.6 Split compound forms (*ifthat, nolde, shalpu*, etc.), retag, etc.
4. For other nodes (detailing deletion, insertion, commentary, etc., also punctuation), figure out an appropriate tag (e.g. COMMENT, INS), reformat slightly, and pass to output file.
5. Use punctuation as (very) rough guide to sentence boundaries, insert IP-MAT brackets accordingly.

So far



```
1. vim
vim vim bash
(IP-MAT | (~f36va~|
(CODE {PAGE~f36va}) | $/P11N_*I^C
(PRO@P11N@ I^c-I) | $see/vSpt11_SEI
(VBD@vSpt11@ sei-see) | $/A0d_A
(D@A0d@ a-a) | $sight/n0d_SIC+TE $-th/xs-n0d_+TE
(N@n0d@ sic=te-sight) | $/RTI_y^T
(C@RTI@ +t^t-that) | $be/vpt13_WAS
(BED@vpt13@ was-be) | $unsee/vSpp_VN-SEIy+E $un-/xp-v_VN- $/v
(VAN@vSpp@ vn=sei+t=e-unsee) | Spp[V]{rh}_+E
(LINEBREAK \) | {\}
(D@A0d@ A-a) | $/A0d_*A
(N@n0d@ find=els-fyndels) | $fyndels/n0d_FIND+ELS $-els/xs-n0d_+ELS
(C@RTI@ +t^t-that) | $/RTI_y^T
(BED@vpt13@ was-be) | $be/vpt13_WAS
(VAN@vSpp{rh}@ vn=fund=e-unfind) | $unfind/vSpp{rh}_VN-FUND+E $un-/xp-v_VN
(PUNC .) | - $/vSpp{rh}_+E
(LINEBREAK \) | {.}
| ) | {\}
westminstert.pos 1,2 Top <ts/westminstert.tag 11,1 6%
```

## Workflow: Parsing

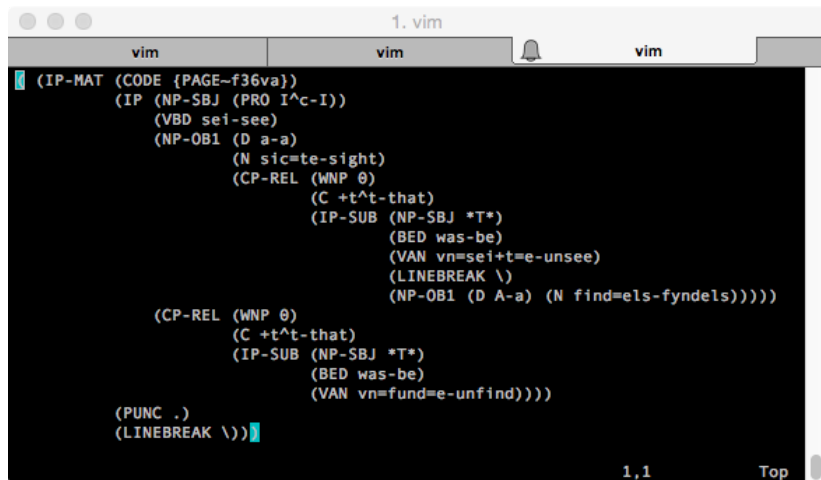
1. Use corpus revision queries to incrementally add/adjust bracketing, relabel, etc.
2. Transfer rhyme information to word-level tags (not yet implemented).
3. Strip out LAEME grammels.
4. Manually correct automatically generated parse.

## Sample corpus revision query

[illegible]



# Automatically generated output



The screenshot shows a vim editor window with a dark background. The title bar at the top reads "1. vim". Below the title bar, there are three tabs, each labeled "vim". The main editing area contains a syntactic tree for the sentence "I saw that he was seeing a cat that was finding a cat". The tree is displayed in a nested, indented format. The root node is (IP-MAT (CODE {PAGE-f36va})). It branches into (IP (NP-SBJ (PRO I^c-I)) (VBD sei-see) (NP-OB1 (D a-a) (N sic=te-sight) (CP-REL (WNP 0) (C +t^t-that) (IP-SUB (NP-SBJ \*T\*) (BED was-be) (VAN vn=sei+t=e-unsee) (LINEBREAK \) (NP-OB1 (D A-a) (N find=els-fyndels)))))))). The (CP-REL (WNP 0) (C +t^t-that) (IP-SUB (NP-SBJ \*T\*) (BED was-be) (VAN vn=fund=e-unfind)))) branch is also present. The tree ends with (PUNC .) and (LINEBREAK \)). The cursor is at the end of the last line.

```
(IP-MAT (CODE {PAGE-f36va}))
  (IP (NP-SBJ (PRO I^c-I))
    (VBD sei-see)
    (NP-OB1 (D a-a)
      (N sic=te-sight)
      (CP-REL (WNP 0)
        (C +t^t-that)
        (IP-SUB (NP-SBJ *T*)
          (BED was-be)
          (VAN vn=sei+t=e-unsee)
          (LINEBREAK \)
          (NP-OB1 (D A-a) (N find=els-fyndels))))))
    (CP-REL (WNP 0)
      (C +t^t-that)
      (IP-SUB (NP-SBJ *T*)
        (BED was-be)
        (VAN vn=fund=e-unfind))))
  (PUNC .)
  (LINEBREAK \))
```

1,1 Top

# Hand-correction: Annotald

Annotald

localhost:8080/rtruswe2

Annotald 1.3.4

Editing: westminstert.psd

Save

Undo

Redo

Idle/Resume

Exit

Tools

Search

Messages

Save success.

Status

Editing.

IP-MAT

CODE (PAGE~f36va)

NP-SBJ

PRO I^c-1

VBD sei-see

NP-OB1

NP

D a-a

N sic=te-sight

CP-REL

WNP-1 0

C +t^t-that

IP-SUB

NP-SBJ \*T\*-1

BED was-be

VAN vn=sei+t=e-unsee

LINEBREAK \

CONJP

NP

D A-a

N find=els-fyndels

CP-REL

WNP-2 0

C +t^t-that

IP-SUB

NP-SBJ \*T\*-2

BED was-be

VAN vn=fund=e-unfind

PUNC .

LINEBREAK \

## Section 4

What next?

# Prospects

- ▶ Initial grant covers  $< \frac{1}{3}$  of LAEME.
- ▶ Longer-term goals: the rest of LAEME, and LAOS.
- ▶ Practical issues:
  - ▶ Up-to-speed annotators can correct 4-500 words/hour, so working through the rest of LAEME and LAOS involves c.1500 hours of someone's time (Jim's?).
  - ▶ Any way to "reuse parses" across parallel texts? Rough calculation: c.390k words of LAEME are versions of texts already parsed in PPCME2/PCMEP or slated for inclusion in our first sample ( $\approx 87\%$  of the remainder of LAEME).
- ▶ Potential for new research uses of parsed historical corpora?
  - ▶ Traditionally, PPCHE people have had limited interest in dialectal variation.
  - ▶ Our initial choice of texts for parsing was motivated by helping PPCHE people to find fuller answers to PPCHE-style questions.
  - ▶ Any hope for using parallel texts in a fully parsed version of LAEME to investigate dialectal variation in syntactic structure?

# References

- Bacovcin, H. A. (2016). Modelling interactions between morphosyntactic changes. In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.
- Ecay, A. & Tamminga, M. (2016). Persistence as a diagnostic of grammatical status: The case of Middle English negation. In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.
- Gisborne, N. & Truswell, R. (2016). Where do relative specifiers come from? In E. Mathieu & R. Truswell (Eds.), *From Micro-change to Macro-change*. Oxford: Oxford University Press.
- Laing, M. (2013). A Linguistic Atlas of Early Middle English, 1150–1325. Version 3.2, <http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html>.